Entity Analytics: Because context matters

Analysts routinely spend up to 80 percent of their time cleaning and validating data prior to building predictive models. This is compounded by the steep challenges of integrating diverse, enterprise-wide data sources. The explosion of available data exacerbates the amount of natural variability. An example is “Bob” versus “Robert,” “IL” versus “Illinois” and “55A” versus “55-A.” Unintentional errors are also likely, such as a reversed month and day in a birthday or transposed part numbers. Data can also be compromised by professionally fabricated lies, such as fake identities or nonexistent product codes. Predictive models that are built with compromised or inaccurate data are by definition modeling an inaccurate universe.

With Entity Analytics, analysts can easily overcome challenges with identity to build better quality models. More accurate models result in better business outcomes, whether the goal is detecting and preempting risk or recognizing and responding to opportunity. At the same time, Entity Analytics can provide better insight into the entities that matter to your business: customers, employees, organizations, vehicles, vessels and more.

SPSS Modeler Entity Analytics adds an extra dimension to SPSS Modeler predictive analytics. Whereas predictive analytics attempts to predict future behavior from past data, entity analytics focuses on improving the coherence and consistency of current data by resolving identity conflicts within the records themselves. An identity can be that of an individual, an organization, an object, or any other entity for which ambiguity might exist. Identity resolution can be vital in a number of fields, including customer relationship management, fraud detection, anti-money laundering, and national and international security.

Suppose that you have the following customer records from two different sources, and are not sure whether they refer to the same person or different people.

Source 1
Record no.: 70001
Name: Jon Smith
Address: 123 Main Street
Tax Reference: 555-00-1111
Driv. License: 0001133107
Cred. Card: 10229127

Source 2
Record no.: 9103
Name: JOHNATHAN Smith
Date of Birth: 06/17/1934
Telephone: 555-1212
Cred. Card: 10229128
Email: jls@mail.com
IP address: 9.50.18.77

There are no exact matches in the data between the two records. However, if we introduce a third source, we find some common attributes.

Source 3
Record no.: 6251
Name: Jon Smith
Telephone: 555-1212
Driv. License: 0001133107
Cred. Card: 10229132

The driving license number links the records in Source 1 and Source 3, while the telephone number links Sources 2 and 3. So we can be reasonably sure that all three sources refer to the same person.

But what if the distinction is not so easy to make? We may have very little data on which to base our decision. Consider the following two records.

Source 4
Record no.: S45286
Name: John T Smith Jr
Address: 456 Main Street
Telephone: 703-555-2000

Date of birth: 03/12/1984

Record no.: S45287
Name: John T Smith
Address: 456 Main Street
Telephone: 703-555-2000
Driv. License: 009900991

Evidently this is not the same Mr Smith from the previous records--the differences are sufficient that we can rule this out. However, we do still have a problem. Two different records, from the same data source, appear to relate to the same person. Are they duplicate records? We cannot be sure unless we can find another related record giving us more information, perhaps from a different source.

Source 5
Record no.: 769582-2
Name: John T Smith Sr
Address: 456 Main Street
Telephone: 703-555-2000
Driv. License: 009900991
Date of birth: 06/25/1959

This resolves the problem. The two records in Source 4 are not duplicates, but are actually a father and son with the same name, living at the same address, and using the same telephone number. On a manual system, it could take weeks of searching to find the one record that resolved the identities. With an automated entity analytics system, resolution time is dramatically reduced.

Entity analytics and predictive analytics

If all of your data consisted of a single source of records that were complete and unambiguous, it would be relatively simple for SPSS Modeler to resolve any identity conflicts. Using only predictive analytics, you could read your data into SPSS Modeler, perform your processing and obtain reliable results.

In the real world, however, the picture is normally very different. Data is typically far from complete, frequently ambiguous, and often scattered over many different data sources, recording many different attributes with few overlapping fields. Part of the value of entity analytics lies in collecting data from all the different sources into a single, central storage area, known as a **repository**. The entity analytics system then examines the data in minute detail to resolve conflicts, adding a unique identifier to records that originate from the same person or organization.

The following table illustrates the differences between the two types of analytics.

Characteristic	Predictive analytics	Entity analytics
Types of training data	Based on relatively small sets and numeric ranges	Can exploit large sets (typeless fields) like names and addresses
Size of training data	Typically ignores large sets (typeless fields)	All data used
Generalization	Algorithm generalizes across training data to form concise model	Data persisted in structures suitable for entity matching and relationship detection

Characteristic	Predictive analytics	Entity analytics
Fraud detection	Records flagged as potentially fraudulent if they have typical characteristics of fraudulent application	Records flagged as potentially fraudulent if related to known fraudulent records, or if originating from same individuals but with different identities

Using entity analytics with SPSS Modeler

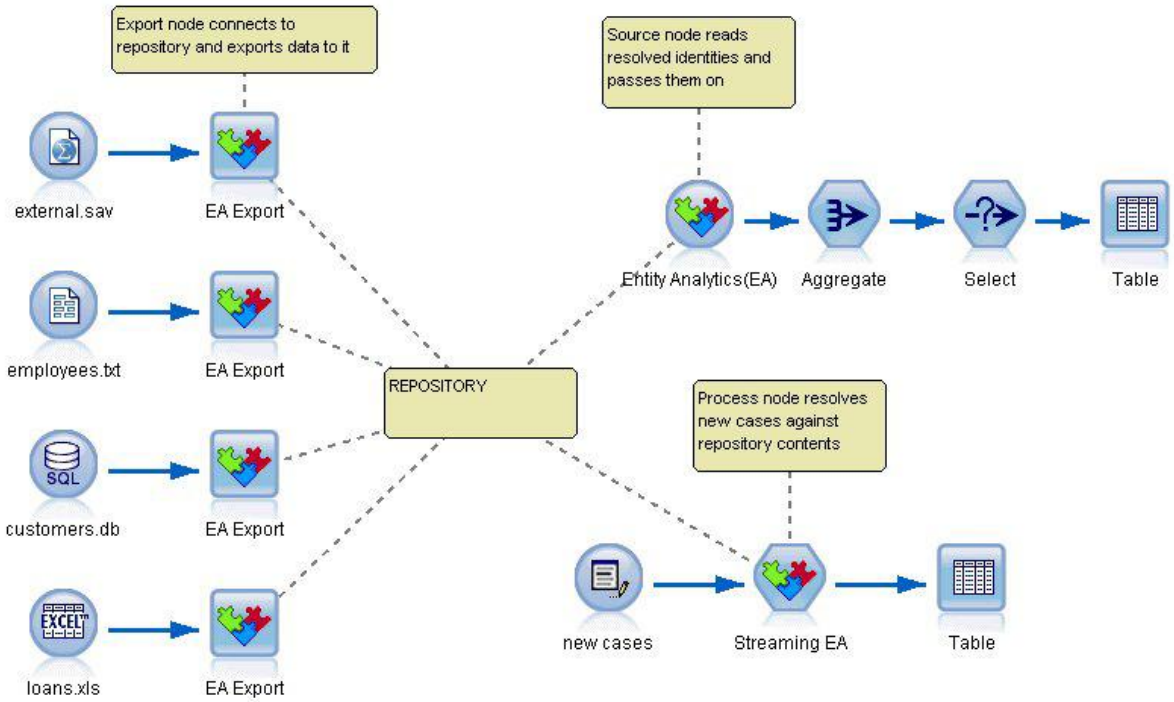
You suspect that you may have identity problems with your data. For example, individuals might appear more than once, or distinct individuals might appear to be merged or missing. How can SPSS Modeler Entity Analytics help you address this? The following is a suggested procedure, though you may need to vary this to suit your particular requirements.

- Read the source data into SPSS Modeler
- Create a repository ready to store the data
- Connect SPSS Modeler to the repository
- Map the data fields to repository features
- Export the data into the repository and resolve the identities
- Analyze the resolved identities
- Resolve new cases against the repository
- Generate any necessary alerts (batch or real-time)

At this point, you need to know something of how SPSS Modeler works. SPSS Modeler is a very user-friendly tool, based on the graphical representation of a stream of data flowing through a number of nodes. Each node represents a particular stage of the workflow.

SPSS Modeler itself provides a wide range of nodes, covering all the standard data mining functions. SPSS Modeler Entity Analytics adds nodes for use specifically in entity analytics. These are EA export node, the Entity Analytics(EA) source node, and the Streaming EA process node.

The following figure illustrates the process.


Export node connects to repository and exports data to it
Source node reads resolved identities and passes them on
external.sav
EA Export
Entity Analytics(EA)
Aggregate
Select
Table
employees.txt
EA Export
REPOSITORY
Process node resolves new cases against repository contents
customers.db
EA Export
new cases
Streaming EA
Table
loans.xls
EA Export